



LAWRENCE
LIVERMORE
NATIONAL
LABORATORY

THE USE OF SINGLE NUCLEOTIDE POLYMORPHISM (SNP) AND MULTIPLE LOCUS VARIABLE NUMBER TANDEM REPEAT (MLVA) ANALYSES TO STUDY THE POPULATION GENETICS OF PATHOGENIC MICROBES

Paul J. Jackson

July 22, 2009

Bacterial Population Genetics in Infectious Disease

Disclaimer

This document was prepared as an account of work sponsored by an agency of the United States government. Neither the United States government nor Lawrence Livermore National Security, LLC, nor any of their employees makes any warranty, expressed or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States government or Lawrence Livermore National Security, LLC. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States government or Lawrence Livermore National Security, LLC, and shall not be used for advertising or product endorsement purposes.

The Use of Single Nucleotide Polymorphism (SNP) and Multiple Locus Variable Number Tandem Repeat (MLVA) Analyses to Study the Population Genetics of Pathogenic Microbes

Paul J. Jackson

Lawrence Livermore National Laboratory

The study of the population genetics of pathogenic microbes has benefited significantly from the increasingly rapid and less costly ability to generate direct DNA sequence information from individual isolates. It is now possible to sequence and assemble a high quality draft sequence of a microbial genome in just a few days. However, circumstances that require analyses of many different microbes or that must be fielded to diagnostic laboratories still rely primarily on methods that interrogate a relatively small sampling of a bacterial genome. Consequently databases of single nucleotide polymorphisms (SNPs) and Multiple Locus Variable Number Tandem Repeats (MLVA) have been generated and are still being populated with SNP and MLVA profiles for different infectious bacterial pathogens. Such databases allow a rapid characterization of a new or unknown isolate relative to all other isolates for which such profiles are available. In the following chapter, the methods used to initially identify SNPs and MLVA profiles will be described, examples of the application of these methods to generate phylogenetic information will be presented, databases in use to allow rapid comparison among isolates will be introduced and limitations of the different methods will be discussed.

Prior to applying any method to differentiate among members of a population of microbial isolates, one must have a sufficiently large and diverse collection of isolates to

ascertain whether the method to be used to differentiate among the isolates will provide sufficient resolution to be useful for further studies. In the absence of a validated method to demonstrate diversity, this can be problematic. In some species diversity has been demonstrated by a number of phenotypic or immunologic methods. In *Yersinia pestis*, the causative agent of bubonic plague, isolates were differentiated by their ability to ferment glycerol and reduce nitrate (Devignat, 1951). In *Bacillus thuringiensis*, a highly diverse insect pathogen of commercial value, diversity is measured by the presence of different insecticidal toxins and diversity in flagellar H-antigen agglutination reactions (Crickmore, *et al.*, 1998, Lecadet and Frachon, 1994). Subspecies differentiation in *F. tularensis*, the causative agent of tularemia was done by biochemical analysis (Johansson, *et al.*, 2000). Biovar A ferments glycerol and glucose and produces citrulline ureidase while biovar B ferments only glucose and does not produce this enzyme (Gurycova, 1998). More recently immunoassays using monoclonal antibodies have been applied to detect and differentiate among different subtypes (Grunow, *et al.*, 2000). Similar methods did not separate different *Bacillus anthracis* isolates into distinctively different groups. Amplified fragment length polymorphism (AFLP) analysis was first used to demonstrate differences among the many isolates with limited diversity within this species (Keim, *et al.*, 1997). Analysis of only the few polymorphic fragments in the AFLP profiles for the different strains analyzed provided the first successful differentiation among multiple isolates of this species. Further analysis of the AFLP profiles revealed several DNA fragments that were mutually exclusive among different isolates. That is, a strain that manifested a particular fragment always lacked a different fragment, often of similar size while a different strain would manifest the second fragment but not the first. These observations led to the discovery of five variable number tandem repeats (VNTRs) in *B. anthracis* (Keim, *et al.*, 2000). However, the

first example of a variable number tandem repeat in *Bacillus anthracis* was discovered by a much less extensive survey of the *B. anthracis* genome using arbitrarily primed (AP)-PCR (Andersen, Simchock and Wilson, 1996). A single genetic locus manifested two different alleles in a limited number of *B. anthracis* isolates. Subsequent analyses using a much larger strain collection (198 isolates) revealed five alleles at this locus (Jackson, *et al.*, 1997). Additional MLVA loci were subsequently identified by first, direct analysis of the available *B. anthracis* pXO1 and pXO2 sequences (Keim, *et al.*, 2000), then by analysis of the entire *B. anthracis* genome (Van Ert, *et al.*, 2007a). MLVA8 analysis, using eight VNTR loci, separates all tested *B. anthracis* isolates into 89 different genotypes. MLVA15 analysis, using 15 different VNTR loci, in combination with a larger number of isolates analyzed increased the genotype number from 89 to 221 (Van Ert, *et al.*, 2007a).

The first VNTR in *Yersinia pestis* was a tetranucleotide repeat, (CAAA)_n where n = 3-10 (Adair, *et al.*, 2000). All possible alleles between three and ten were found in a survey of 35 diverse *Y. pestis* isolates. The (CAAA)₂ was found in *Yersinia pseudotuberculosis*, a close relative of *Y. pestis* but the same VNTR was not found in *Y. enterocolitica*, another relatively close relative. Analysis of *Y. pestis* chromosomal DNA sequences and the sequences of two plasmids, pMT1 and pCD1, identified an additional 42 VNTR loci in *Y. pestis* (Klevytska, *et al.*, 2001). VNTR-based phylogenetic trees were generally consistent with common biovar evolutionary scenarios and with IS100-based analyses (Motin, *et al.*, 2002). Pourcel, *et al.* (2004) used 25 *Y. pestis* MLVA markers to characterize 180 different isolates into 61 different genotypes. The three traditional *Y. pestis* biovars consistently distributed into three branches with some exceptions, primarily in the Medievalis biovar. Studies of VNTR mutation rates in *Y. pestis* and other species allow application of this technology to better epidemiological

understanding of disease outbreaks and their progression (Vogler, *et al.*, 2007).

Multiple locus variable number tandem repeat analyses have also been applied to distinguish among different *Escherichia coli* O157:H7 isolates. Twenty-nine putative VNTR loci were identified by interrogation of the *E. coli* genome and these were validated by analyses of 56 different *E. coli* O157:H7/HN and O55:H7 isolates (Keys, Kemper and Keim, 2005). The number of alleles at each locus ranged from two to 29 while the diversity index varied from 0.23 to 0.95. Values of this index can range from 0 (no diversity) to 1 (complete diversity) (Nei and Kumar, 2000). A comparison of MLVA typing to Pulsed Field Gel Electrophoresis (PFGE) results showed that both methods provided consistent results but MLVA analyses were able to further resolve among sample isolates that were identical by PFGE analysis. Thus, MLVA analysis of an outbreak cluster should generate superior resolution to the more traditional PFGE methods in addition to being somewhat easier and significantly more rapid to execute. MLVA analysis was used to better understand *E. coli* O157:H7 contamination of lettuce and spinach in the Salinas and San Juan valleys of California between 1995 and 2006 (Cooley, *et al.*, 2007). A comparison to PFGE results again demonstrated resolution among apparently identical isolates by MLVA analysis. MLVA analysis of fifty-four feedlot isolates separated into 12 different MLVA types and suggested that animals entering the feedlot at initial stocking are an important source of this contamination. Once *E. coli* O157 inoculated the feedlot, water troughs, pen bars, pen floor feces and feed were all found to be means of transmitting this pathogen (Murphy, *et al.*, 2008). Seventy-two human and animal strains of Shiga-toxin-producing *E. coli* O157 were typed using MLVA assays (Lindstedt, *et al.*, 2003); and Ohata *et al.* (2008) typed Japanese *E. coli* O157:H7 clinical isolates. Comparisons in both studies again demonstrated the superior resolution of MLVA typing relative to PFGE analysis.

Francisella tularensis, the etiologic agent of tularemia is found naturally throughout the northern hemisphere in North America, Asia and Europe although it has also been isolated in Australia. This highly infective, gram-negative intracellular pathogen can infect a large number of different species. There are four subspecies of *F. tularensis*. The subspecies *tularensis* is the most virulent of these. Biochemical studies and 16S rDNA sequence analysis have traditionally been used to distinguish among different subspecies of this pathogen. Six polymorphic VNTR were initially identified based on canvassing the *F. tularensis* genome followed by analysis of these putative VNTR loci in 55 different *F. tularensis* isolates (Farlow, *et al.*, 2001). The allele number in these six loci ranged from two to 20. Analysis of an additional 56 samples resulted in identification of 39 different allele combinations. UPGMA cluster analysis revealed two major clusters of isolates. However, there were no absolute fixed allelic differences between the two clusters. California isolates were found in both major groups. Oklahoma isolates mapped to one of two sub-groups within the second cluster while all of the Arizona isolates analyzed appeared to be identical at the resolution of the MLVA analysis. An additional 19 variable VNTR loci containing between 2 and 31 alleles were used to analyze 192 geographically diverse *F. tularensis* isolates (Johansson, *et al.*, 2004). Nei's diversity values ranged between 0.05 and 0.95 and were correlated with the number of alleles at each locus. *F. tularensis* subsp. *tularensis* (type A) isolates showed great diversity but *F. tularensis* subsp. *holarctica* (type B) isolates were much less diverse in spite of a much broader geographic range. Some but not all genetically similar isolates were isolated from geographically proximal locations.

Burkholderia pseudomallei is a genetically diverse pathogen that causes a disease called melioidosis. This disease is endemic throughout Southeast Asia and Northern Australia (White, 2003; Cheng and Currie, 2005) and the number of cases increases significantly during the wet

monsoon season. PFGE, AFLP and multi-locus sequence typing (MLST, see below) have been used to distinguish among different isolates of this pathogen. *B. pseudomallei* has numerous VNTRs, some duplicated at more than one site within the genome. Duplicated repeat regions may facilitate genomic rearrangement and, possibly, altered gene expression. However, they can significantly complicate a MLVA analysis and are therefore not usually used in such analyses. U'Ren, *et al.*, (2007) used 32 VNTR loci displaying between 7 and 28 alleles, with Nei's diversity values ranging between 0.47 and 0.94 to analyze 66 geographically diverse *B. pseudomallei* and 21 *Burkholderia mallei* isolates. They also applied these assays to 95 lineages of an 18,000 generation passage experiment to better understand the mutation frequencies at the different VNTR loci. MLVA-based phylogenetic analyses of the *Burkholderia* isolates demonstrated that the *B. mallei* isolates were significantly less diverse, clustered tightly relative to all of the *B. pseudomallei* isolates. Similar results were demonstrated using AFLP and MLST analyses. Analysis of isolates from the passage experiment revealed that variation in 12 of the VNTR loci occurred during the passage study. Most of these changes resulted in single repeat changes with a bias towards increases in tandem repeat copy number. More recently, Currie, *et al* (2009) developed a four-locus MLVA analysis for rapid typing of *B. pseudomallei* based on selection of a sub-set of informative VNTR markers. This analysis provides resolution similar to PFGE and MLST results and can provide genotyping results within 8 hours following receipt of samples.

There are a significant number of other pathogenic microbes besides those outlined here that have been analyzed using the MLVA approach. The utility of a technique that could separate isolates into more than two categories by analysis of a single locus is obvious. Studies demonstrate that the “mutation frequency” of VNTRs is significantly higher than that of other

types of mutations suggesting that MLVA analyses provide greater resolution than analysis of most other changes while interrogating fewer loci. On a per locus basis, VNTRs generally contain greater discriminatory capacity than any other type of molecular typing system (Richards and Sutherland, 1997; van Belkum, *et al.*, 1998). Most pathogenic microbes contain VNTRs but, in some highly diverse species, not all isolates contain the same complement of these loci.

There are online databases for those who wish to conduct MLVA analyses of different microbes (Grissa, *et al.*, 2008). One such site can be found at <http://minisatellites.u-psud.fr/>. The Institut Pasteur maintains a MLVA database that can be accessed at www.pasteur.fr/mlva. However, MLVA profiles are available for only a small number of pathogens. Often such databases are limited to species of interest to those who constructed the site. The author could not find a single online source containing all MLVA profiles for all the pathogens mentioned in this chapter.

Perhaps the most difficult aspect of using MLVA analysis is the difficulty of comparing MLVA analyses from different laboratories. Early MLVA analyses were often conducted using polyacrylamide-based DNA sequencing gels. Such gels and the current capillary-based electrophoresis instruments were designed to resolve single-nucleotide differences between DNA fragments, not to determine specific fragment sizes. Calling the size (in bp) of a fragment on a gel-based sequencer is within ± 3 nucleotides depending on the number of molecular weight markers included in the analysis. MLVA fragment lengths may differ by less than three nucleotides. Moreover, commercially available DNA fragment sizing standards provide different putative lengths because DNA standard fragments migrate in the gel based on both the length of the sequences and their nucleotide content. That is, two fragments of exactly the same nucleotide length may migrate slightly differently on the gel, resulting in a different fragment

length call for a MLVA allele relative to the same analysis with a different set of size standards. Consequently fragment length calls can differ significantly among different laboratories. This problem is magnified when using capillary gel electrophoresis, where DNA fragment length calls vary ± 9 nucleotides or more. Direct measurement of fragment masses by mass spectrometry can provide very accurate analysis of MLVA results, but such instruments are very expensive and not readily available to most laboratories conducting such analyses. One approach to solving this problem is to make DNA fragment size standards directly from a set of different MLVA alleles at a particular VNTR locus. This allows direct comparisons that can be shared among laboratories using the same fragment array for comparison. For example, the *vrpA* VNTR locus of *B. anthracis* has five alleles, containing from two to six repeats, 12 nucleotides in length. Generation of a set of all five alleles differentially labeled relative to the assay, allows direct comparison of an assay result to an array of all five possible alleles (Figure 1). This provides a relatively simple determination of the allele present in an unknown sample. However, direct comparison of alleles among laboratories requires that all laboratories use the same sets of size standards; that is, a set of size standards for each MLVA locus that contains fragments representing all of the alleles at that locus. Such specialized molecular weight markers are not generally available. One can also identify MLVA loci and the number of repeats present by analysis of complete or partial genome sequences from different bacterial isolates (Denœud and Vergnaud, 2004). This, of course, assumes that the sequences are already available because the cost and time of sequencing and assembling a genome are still significantly more than MLVA analyses with developed assays and reagents.

Development of alternative DNA-based methods of sample analysis were initially driven by the cost and time required to generate high quality DNA sequences and the increased use of

single nucleotide polymorphisms (SNPs) to differentiate among different microbial isolates has closely paralleled the increased access to low cost DNA sequencing. Recently, introduction of new, much more rapid methods of generating DNA sequences, especially for comparison to previously sequenced organisms has made the initial identification of SNPs across multiple isolates of the same species and among multiple, closely related species much more rapid and less expensive than was previously possible. Indeed, potential VNTR loci are now almost exclusively identified by analysis of microbial genomes (Dégrange, S. *et al.*, 2009 for a recent example).

Historically, the first extensively used SNP-based method of characterizing microbes was analysis of a specific region of the 16S ribosomal RNA gene common to all microbes (Lane, *et al.*, 1985). The sequence of the small-subunit (16S) ribosomal RNA (rRNA) varies in an orderly manner across phylogenetic lines and contains segments that are conserved at the species, genus, or kingdom level. By designing oligonucleotide primers to prime off sequences conserved throughout the eubacterial kingdom, it is possible to use PCR to amplify DNA fragments encoding phylogenetically informative sections of the 16S RNA gene. Subsequent sequencing of these amplicons provides information that is useful to identify an isolate at least to the genus level. Sometimes this approach provides differentiation among different isolates of highly diverse species (Collins and East, 1998) although, based on the 16S rRNA results, one could argue that the species in question is actually multiple species. The 16S rRNA typing approach is still widely used to generate information about previously uncharacterized isolates. The Ribosomal Database Project (<http://rdp.cme.msu.edu/>), archives over 920,000 16S rRNA sequences with the software and informatics tools for comparison to sequences generated from unknown isolates.

The 16S rRNA sequences vary little among some closely related species. For example, this approach will not differentiate between *B. anthracis* and some closely related *B. cereus* isolates when amplifying the template normally generated when analyzing unknown *Bacillus* isolates. Another method, analyzing selected sequences of highly conserved so-called “housekeeping genes” provides higher resolution among different closely related microbial isolates. Multi-locus sequence typing (MLST) was first applied to analyze populations of pathogenic microbes by Maiden, *et al.* in 1998 in a study of *Neisseria meningitidis*. The study analyzed fragments approximately 470 nucleotides in length from 11 different genes. The amplicon size was selected to provide rapid full sequencing of both amplicon DNA strands using the best available sequencing technology of the time. Most MLST analyses now target portions of seven different conserved genes. Analysis of the *Bacillus cereus* group relies on sequencing portions of the *glpF*, *gmk*, *ilvD*, *pta*, *pur*, *pycA*, and *tpi* genes, encoding the glycerol uptake facilitator protein, guanylate kinase, dihydroxy-acid dehydratase, phosphate acetyltransferase, phosphoribosylaminoimidazolecarboxamide, pyruvate carboxylase, and triosephosphate isomerase respectively. Hoffmaster, *et al.* (2006) applied MLST and AFLP analysis to *B. cereus* isolates associated with fatal pneumonias. The results were complementary. Isolates that appeared to be closely related by AFLP analysis were also closely linked by MLST analysis. Isolates that appeared to be identical by MLST analysis were also identical within the resolution of the AFLP analysis. Both methods allowed comparison across a large, diverse collection of *B. cereus* and *B. thuringiensis* isolates. Neither method provided significant resolution among different *B. anthracis* isolates but clearly differentiated all *B. anthracis* isolates from even very closely related *B. cereus* isolates.

Y. pestis and its closest relatives *Y. pseudotuberculosis* and *Y. enterocolitica* have been

subjected to MLST analyses by sequencing portions of the *thrA*, *trpE*, *glnA*, *tmk*, *dmsA*, and *manB* genes (Achtman *et al.*, 1999). The MLST results in combination with other information about the genome organization in different isolates of these species support the contention that *Y. pestis* is a recently emerged clone on *Y. pseudotuberculosis*. Another study of *Y. pestis* isolates from the Republic of Georgia and neighboring former Soviet Union countries applied MLST to differentiate among different isolates (Revazishvile, *et al.*, 2008). However, analysis at seven loci (portions of the *hsp60*, *glnA*, *gyrB*, *recA*, *manB*, *thrA* and *tmk* genes) and the 16S rRNA gene provided little resolution and the authors found that PFGE discriminated among the *Y. pestis* isolates more effectively than MLST. It appears that, like *B. anthracis*, *Y. pestis* isolates show little diversity based on such analyses. The lack of diversity within the MLST loci analyzed is in direct contrast to methods that interrogate changes in genome organization, suggesting that many differences among different *Y. pestis* isolates may be based on differences in the relative spatial distribution of sequences within the genome (Motin, *et al.*, 2002).

MLST-based methods have also been applied to the study of *E. coli* isolates. Noller, *et al.* (2003) found no sequence diversity in the sequenced portions of seven housekeeping genes among 77 *E. coli* O157:H7 isolates shown to be diverse using PFGE. Reid, *et al.* (2000) also looked at seven housekeeping genes in *E. coli* strains representing common clones of enteropathogenic *E. coli* (EPEC), an important cause of infantile diarrhea; enterohaemorrhagic *E. coli* (EHEC), one of the primary food-borne pathogens in the industrialized world; strains of other Shiga-toxin-producing *E. coli* serotypes and the laboratory strain K-12 in an attempt to better understand the evolution of new bacterial pathogens. While MLST analyses provided useful data to draw conclusions about evolutionary mechanisms, the diversity index within the fragments of the housekeeping genes amplified ranged from approximately 2% in the *arcA* gene

to only 7.5% in the *mtlD* gene, again suggesting limited value of the MLST approach to differentiate among even very different *E. coli* isolates.

MLST has also been applied to *F. tularensis* and closely related species. However, Johansson, Forsman, and Sjöstedt (2004) showed that the use of seven housekeeping genes of *F. tularensis* distinguished the subspecies but did not provide high-resolution discrimination of individual isolates. In contrast, MLVA analysis was only exceeded by whole genome sequencing in providing resolution among different *F. tularensis* isolates.

There are several published reports describing application of MLST to the study of *Burkholderia pseudomallei* and *B. mallei* isolates (Godoy, *et al.*, 2003; Currie, *et al.*, 2007; Wattiau, *et al.*, 2007). MLST analysis of 128 isolates of a geographically diverse collection of *B. pseudomallei* isolates using sequences from the *ace*, *gltB*, *gmhD*, *lepA*, *lipA*, *narK* and *ndh* genes resolved the collection into 71 sequence types (Godoy, *et al.*, 2003). Resolution was improved by the presence of multiple SNPs within the different amplicons. For example, there were 15 different SNPs in the *gmhD* gene fragment and 14 SNPs in the *narK* gene fragment. Specific nucleotides present at five different SNP loci can be used together to unambiguously differentiate between all *B. mallei* and all *B. psuedomallei* isolates tested (Okinaka, R., unpublished).

Commercially available SNP analysis kits and custom synthesized primers and probes are now routinely available. As outlined above, SNP analyses can be applied, with varying success that depends on the species in question, to differentiate among different isolates of the same species or across a group of closely related species depending on the targets chosen for the analysis.

Demonstration of the resolution of any typing method requires signatures for a large

collection of diverse isolates and this is sometimes the limiting factor in demonstrating the utility of such methods, especially when analyzing species where there is a lack of genetic diversity among the isolates. There are a significant number of publications describing application of MLST methods to differentiate among different, closely related species or among different isolates of the same species. However, another inherent weakness in this approach is the lack of common databases containing all of the MLST profiles for a particular target species. MLST profiles for the *Bacillus cereus* group and *Burkholderia pseudomallei* (also applicable to *B. mallei*) are available at PubMLST (<http://pubmlst.org/>) but the database contains a relatively small number of isolates relative to the large collections available. While *Y. pestis* and its closest relatives *Y. pseudotuberculosis* and *Y. enterocolitica* have been subjected to MLST analyses by sequencing portions of the *thrA*, *trpE*, *glnA*, *tmk*, *dmsA*, and *manB* genes (Achtman *et al.*, 1999), the available database for MLST analysis of *Y. tuberculosis* (<http://mlst.ucc.ie/mlst/dbs/Ypseudotuberculosis>) provides profiles for a somewhat different set of genes; *thrA*, *trpE*, *glnA*, *tmk*, *adk*, *argA*, and *aroA*. Two different *E. coli* MLST databases are available (<http://www.pasteur.fr/recherche/genopole/PF8/mlst/EColi.html> and <http://mlst.ucc.ie/mlst/dbs/Ecoli>) but they each provide profiles for a different set of target genes with only one common target. Besides containing profiles for a limited number of profiles, the lack of consensus on which genes to target is a significant weakness. Analysis of mutually exclusive sets of target genes does not allow comparison of results. Perhaps a forum could be established to determine which MLST loci provide the highest resolution among all available isolates for particular species. It would clearly be beneficial if consensus on the genes targeted for analysis was reached.

There is also a website for comparison of *Burkholderia pseudomallei* MLST profiles and those from closely related species that targets seven MLST loci (www.bpseudomallei.mlst.net/earth/map/). This website provides information about the geographic location (by country) of a large collection of isolates as well as the MLST genotype for the collection of isolates from that country.

MLST analysis is advantageous because it exploits the unambiguous nature of DNA sequences to allow data comparison across multiple laboratories. However, in some species, it provides only very limited resolution among demonstrated diverse isolates. It is also dangerous to make general assumptions about evolutionary and phylogenetic differences based only on analysis of changes in a limited number of gene sequences. Sequence differences identify one kind of diversity but do not capture whole genome changes such as the relative presence or absence of particular sequences or genome reorganization among closely related species or among different isolates of the same species. It has been clearly demonstrated in *Y. pestis* that genome reorganization, probably facilitated by the large number of IS elements present, contributes to phenotype among different isolates of this pathogen and, perhaps, in differentiating between this pathogen and other closely related bacterial species.

Another approach to differentiating among isolates of the same pathogenic species focuses on changes within genes that are unique to that species. In particular, studies have targeted genes encoding known virulence or toxin factors. Price, *et al.* (1999) showed that *B. anthracis* isolates could be differentiated by analysis of the protective antigen gene. Analysis of seven SNPs within this gene in combination with MLVA analysis of different isolates provided a unique signature for an isolate associated with the 1979 Sverdlovsk release. Twelve SNPs have now been identified in the protective antigen gene (Jackson, P.J., unpublished). Five additional

SNPs have been found in the *cya* gene, encoding edema factor also residing, with the protective antigen gene, on pXO1 (Okinaka, R.T., unpublished). Combinations of SNPs in a single gene have been used to differentiate between a target species and its close relatives. Qi, *et al.*, (2001) identified four SNPs within the *rpoB* gene that were reported to be specific for *B. anthracis* relative to other *Bacillus* isolates. However, comparisons were not made between *B. anthracis* and *B. cereus* and *B. thuringiensis* isolates that have been shown by other methods to be very closely related to this pathogen. In the absence of either a discrete genetic change that correlates with the SNP – for example, the expression of a particular gene characteristic of the pathogen – or a very extensive survey of closely related species, it is dangerous to assume that one or a very limited number of SNPs may, in themselves, represent a species-specific assay.

The availability to fully sequenced genomes from different genetically diverse isolates of the same pathogenic species has led to extensive surveys that identify virtually all of the phylogenetically informative SNPs in a genome. Comparative full-genome sequencing among eight *B. anthracis* strains led to discovery of approximately 3,500 SNPs (Read, *et al.*, 2002; Pearson, *et al.*, 2004). It would seem that the binary nature of SNPs provide only limited subtyping power with a large number of SNPs required to provide the resolution needed to differentiate among closely related isolates. However, it has been shown that a surprisingly small number of SNPs can be used to provide high definition resolution among different genetic groups (Van Ert, *et al.*, 2007b). Keim *et al.* (2004) developed this concept further and proposed the “canonical SNP,” a SNP that can be used to define a point in the evolutionary history of a species. Such “canonical SNPs” can be used diagnostically to define major genetic lineages within a species or, more narrowly, to define specific isolates. Moreover, combining “canonical SNP” and MLVA analyses provides insights into the evolutionary history of the species. A set

of only 12 “canonical SNPs” representing different points in the evolutionary history of *B. anthracis* were used, in combination with MLVA15 analyses to type a large, diverse, global collection of *B. anthracis* isolates. SNP analyses placed all isolates into 12 conserved groups or lineages (Van Ert, *et al.*, 2007b). The analysis of the slowly evolving “canonical SNP” in combination with the MLVA15 results greatly enhanced the resolution beyond the 221 genotypes resolved by MLVA15 analysis alone. Analysis of slowly evolving “canonical SNPs” allowed definition of major clonal lineages while younger, population-level structure was revealed using the more rapidly evolving MLVA markers.

SNP analyses can also detect changes of a more sinister nature. Resistance to ciprofloxacin in *B. anthracis* results from a number of single nucleotide changes in the *gyrA* and *parC* genes, encoding the proteins targeted by this antibiotic (Price, *et al.* 2003). It is highly unlikely that naturally occurring *B. anthracis* isolates will be resistant to this antibiotic because anthrax is primarily a zoonotic disease and infected animals are seldom provided antibiotic therapy to treat their condition. Therefore, the presence of one or more SNPs in critical positions in these genes suggests that an isolate containing such SNPs may have been intentionally subjected to increasing antibiotic concentrations to select a ciprofloxacin-resistant *B. anthracis* isolate. In particular, initial selection for ciprofloxacin resistance results in a high frequency of mutations at only two specific nucleotides (Price, *et al.*, 2003; Jackson, P.J., unpublished), allowing rapid screening for such isolates with just two assays. It will likely be possible to identify and develop assays for other such phenotypic changes with the increased ease of producing whole genome sequences and comparing these to similar isolates with slightly different phenotypes.

The availability of multiple whole genome sequences also allows design of high-density microarrays that can be used to compare different isolates of the same species or closely related microbial species (Zwick, *et al.* 2008). In principle, microarrays should allow rapid identification of even minor differences between the array sequence and the challenge DNA. However, the relatively high frequency of “false positive” SNPs exhibited by microarray data does not allow efficient identification of very minor differences relative to the arrayed sequence. Array technology can be used to demonstrate total, additive differences between a reference genome of a particular isolate and those of other isolates. It can also rapidly screen a large number of isolates to provide information about the relationship of an unknown isolate relative to a reference. When the array hybridization results are compared to the most recent available genome sequences for the same strains, the arrays correctly identify strains 100% of the time (K. Gardner, *et al.*, 2009).

The development of different DNA-based methods to interrogate and distinguish among different species and strains of pathogenic microbes has roughly paralleled development of more rapid, less expensive DNA sequencing technologies. In the absence of easily and rapidly obtained whole genome sequences from multiple isolates of the same species, methods that indirectly detected differences among different species and different isolates of the same species were developed. Early assays were based on methods that used restriction endonucleases and basic PCR methods to demonstrate differences among species or isolates. These included AFLP, single VNTR, then, later, MLVA analyses. As the cost of generating DNA sequences continued to drop and the speed with which sequences could be generated increased, assay methods that could exploit the newly available direct sequence information were developed. These involved, first, single SNP assays then, multiple SNP assays and, finally, as the significance of specific

SNPs became more apparent with increased information from multiple isolates of the same species, multiple SNP assays developed to interrogate genetically or phylogenetically significant changes.

SNPs represent evolutionarily slow genome changes relative to changes in variable number tandem repeat loci. Approaches that use a combination of SNP and MLVA analyses can therefore provide significant insights into definition of major clonal lineages and population-level structure within a species.

Prepared by LLNL under Contract DE-AC52-07NA27344

References.

- Achtman, M., Zurth, K., Morelli, G., *et al.* (1999) *Yersinia pestis*, the cause of plague, is a recently emerged clone of *Yersinia pseudotuberculosis*. *Proceedings of the National Academy of Sciences of the United States of America*, **96**, 14033-14048.
- Adair, D.M., Worsham, P.L., Hill, K.K., *et al.* (2000) Diversity in a variable-number tandem repeat from *Yersinia pestis*. *Journal of Clinical Microbiology*, **38**, 1516-1519.
- Andersen, G.L., Simchock, J.M., and Wilson, K.H. (1996) Identification of a region of genetic variability among *Bacillus anthracis* strains and related species. *Journal of Bacteriology*, **178**, 377-384.
- Cheng, A.C., and Currie, B.J. (2005) Melioidosis: epidemiology, pathophysiology and management. *Clinical Microbiology Reviews*, **18**, 383-416.
- Collins, M.D., and East, A.K. (1998) Phylogeny and taxonomy of the food-borne pathogen *Clostridium botulinum* and its neurotoxins. *Journal of Applied Microbiology*, **84**, 5-17.
- Cooley, M., Carychao, D., Crawford-Miksza, L., *et al.* (2007) Incidence and tracking of *Escherichia coli* O157:H7 in a major produce production region in California. *PloS One*, **2**, e1159.
- Crickmore, N., Zigler, D.R., Feitelson, *et al.* (1998) Revision of the nomenclature for the *Bacillus thuringiensis* pesticidal crystal proteins. *Microbiology and Molecular Biology Reviews*, **62**, 807-813.
- Currie, B.J., Haslem, A., Pearson, T., *et al.* (2009) Identification of melioidosis outbreak by multilocus variable number tandem repeat analysis. *Emerging Infectious Diseases*, **15**, 169-174.
- Currie, B.J., Thomas, A.D., Godoy, D., *et al.*, (2007) Australian and Thai isolates of *Burkholderia pseudomallei* are distinct by multilocus sequence typing: revision of a case of

mistaken identity. *Journal of Clinical Microbiology*, **45**, 3828-2829.

Dégrange, S., Cazanave, C., Charron, A., *et al.* (2009) Development of Multiple-Locus Variable-Number Tandem-Repeat Analysis for Molecular Typing of *Mycoplasma pneumoniae*. *Journal of Clinical Microbiology*, **47**, 914-923.

Denc ud, F. and Vergnaud, G. (2004) Identification of polymorphic tandem repeats by direct comparison of genome sequence from different bacterial strains: a web-based resource. *BMC Bioinformatics*, **5**, 4.

Devignat, R. (1951) Varietes de l'espece *Pasteurella pestis*. Nouvelle hypothese. *Bulletin World Health Organization*, **4**, 247–263.

Farlow, J., Smith, K.L., Wong, J. *et al.* (2001) *Francisella tularensis* strain typing using multiple-locus, variable number tandem repeat analysis. *Journal of Clinical Microbiology*, **39**, 3186-3192.

Gardner, S.N., McLoughlin, K.S., Jaing, C.J., *et al.*, (2009) High-Throughput Genomic Polymorphism Analysis of 12 Bacterial Pathogens, In review.

Godoy, D., Randle, G., Simpson, A.J., *et al.* (2003) Multilocus Sequence Typing and Evolutionary Relationships among the Causative Agents of Melioidosis and Glanders, *Burkholderia pseudomallei* and *Burkholderia mallei*. *Journal of Clinical Microbiology*, **41**, 2068-2079.

Grissa, I., Bouchon, P., Pourcela, C., and Vergnaud, G. (2008) On-line resources for bacterial micro-evolution studies using MLVA or CRISPR typing. *Biochimie*, **90**, 660-668.

Grunow, R., Splettstoesser, W., McDonald, S., *et al.* (2000) Detection of *Francisella tularensis* in biological specimens using a capture enzyme-linked immunosorbent assay, an immunochromatographic handheld assay, and a PCR. *Clinical and Diagnostic Laboratory*

Immunology, **7**, 86–90.

Gurycova, D. (1998) First isolation of *Francisella tularensis* subsp. *Tularensis* in Europe. *European Journal of Epidemiology*, **14**, 797-802.

Hoffmaster, A.R., Hill, K.K., Gee, J.E., *et al.* (2006) Characterization of *Bacillus cereus* isolates associated with fatal pneumonias: strains are closely related to *Bacillus anthracis* and harbor *B. anthracis* virulence genes. *Journal of Clinical Microbiology*, **44**, 3352-3360.

Jackson, P.J., Walthers, E.A., Kalif, A.S., *et al.* (1997) Characterization of the variable-number tandem repeats in *vrnA* from different *Bacillus anthracis* isolates. *Applied and Environmental Microbiology*, **63**, 1400-1405.

Johansson, A., Berglund, L., Eriksson, U., *et al.* (2000) Comparative analysis of PCR versus culture for diagnosis of ulceroglandular tularemia. *Journal of Clinical Microbiology*, **38**, 22–26.

Johansson, A., Farlow, J., Larsson, P., *et al.* (2004) Worldwide genetic relationships among *Francisella tularensis* isolates determined by multiple-locus variable-number tandem repeat analysis. *Journal of Bacteriology*, **186**, 5808-5818.

Johansson, A., Forsman, M., and Sjöstedt (2004) The development of tools for diagnosis of tularemia and typing of *Francisella tularensis*. *Acta Pathologica Microbiologica et Immunologica Scandinavica*, **112**, 898-907.

Keim, P., Kalif, A., Schupp, J., *et al.* (1997) Molecular Evolution and diversity in *Bacillus anthracis* as detected by amplified fragment length polymorphism markers. *Journal of Bacteriology*, **179**, 818-824.

Keim, P., Price, L.B., Klevytska, *et al.* (2000) Multiple-locus variable-number tandem repeat analysis reveals genetic relationships within *Bacillus anthracis*. *Journal of Bacteriology*, **182**,

2928-2936.

Keim, P., Van Ert, M.N., Pearson, T., *et al.*, (2004) Anthrax molecular epidemiology and forensics: using the appropriate marker for different evolutionary scales. *Infection, Genetics and Evolution*, **4**, 205-213.

Keys, C., Kemper, S., and Keim, P. (2005) Highly diverse variable number tandem repeat loci in *E. coli* O157:H7 and O55:H7 genomes for high-resolution molecular typing. *Journal of Applied Microbiology*, **98**, 928-940.

Klevytska, A.M., Price, L.B., Schupp, J.M. *et al.* (2001) Identification and characterization of variable number tandem repeats in the *Yersinia pestis* genome. *Journal of Clinical Microbiology*, **39**, 3179-3185.

Lane, D.J., Pace, B., Olsen, G.J. *et al.* (1985) Rapid determination of 16S ribosomal RNA sequences for phylogenetic analyses. *Proceedings of the National Academy of Sciences of the United States of America*, **82**, 6955-6959.

Lecadet, M.-M., and Frachon, E. (1994) Presented at the XXVIIth Annual Meeting of the Society for Invertebrate Pathology, Montpellier, France.

Lindstedt, B.-A., Heir, E., Gjernes, E., Vardund, T., and Kapperud, G. (2003) DNA fingerprinting of shiga-toxin producing *Escherichia coli* O157 based on Multiple-Locus Variable-Number Tandem-Repeats Analysis (MLVA). *Annals of Clinical Microbiology and Antimicrobials*, **2**, 12.

Maiden, M.C., Bygraves, J.A., Feil, E. *et al.* (1998) Multilocus sequence typing: A portable approach to the identification of clones within populations of pathogenic microorganisms. *Proceedings of the National Academy of Sciences of the United States of America*, **95**, 3140-3145.

- Motin, V.L., Georgescu, A.M., Elliott, J.M., *et al.* (2002) Genetic variability of *Yersinia pestis* isolates as predicted by PCR-based IS100 genotyping and analysis of structural genes encoding Glycerol-3-Phosphate and Dehydrogenase (*glpD*). *Journal of Bacteriology*, **184**, 1019-1027.
- Murphy, M., Minihan, D., Buckley, J.F., *et al.* (2008) Multiple-locus variable number of tandem repeat analysis (MLVA) of Irish verocytotoxigenic *Escherichia coli* O157 from feedlot cattle: uncovering strain dissemination routes. *BMC Veterinary Research*, **4**, 2.
- Nei, M. and Kumar, S. (2000) *Molecular Evolution and Phylogenetics*. Oxford University Press, New York.
- Noller, A.C., McEllistrem, M.C., Stine, O.C., *et al.* (2003) Multilocus sequence typing reveals a lack of diversity among *Escherichia coli* O157:H7 isolates that are distinct by pulsed-field gel electrophoresis. *Journal of Clinical Microbiology*, **41**, 675-679.
- Ohata, K., Sugiyama, K., Masuda, T. *et al.* (2008) Molecular typing of Japanese *Escherichia coli* O157:H7 isolates from clinical specimens by multilocus variable-number tandem repeat analysis and PFGE. *Journal of Medical Microbiology*, **57**, 58-63.
- Pearson, T., Busch, J.D., Ravel, J. *et al.* (2004) Phylogenetic discovery bias in *Bacillus anthracis* using single-nucleotide polymorphisms from whole genome sequencing. *Proceedings of the National Academy of Sciences of the United States of America*, **101**, 13536–13541.
- Pourcel, C., André-Mazeaud, F., Neubauer, H., Ramisse, F., and Vergnaud, G. (2004) Tandem repeats analysis for the high resolution phylogenetic analysis of *Yersinia pestis*. *BMC Microbiology*, **4**, 22.
- Price, L.B., Hugh-Jones, M., Jackson, P.J., and Keim, P. (1999) Genetic diversity in the protective antigen gene of *Bacillus anthracis*. *Journal of Bacteriology*, **181**, 2358-2362.
- Price, L.B., Vogler, A., Pearson, T., *et al.* (2003) In vitro selection and characterization of

Bacillus anthracis mutants with high-level resistance to ciprofloxacin. *Antimicrobial Agents and Chemotherapy*, **47**, 2362-2365.

Qi, Y., Patra, G., Liang, X., *et al.* (2001) Utilization of the *rpoB* gene as a specific chromosomal marker for real-time PCR detection of *Bacillus anthracis*. *Applied and Environmental Microbiology*, **67**, 3720-3727.

Read, T.D., Salzberg, S.L., Pop, M., *et al.* (2002) Comparative genome sequencing for discovery of novel polymorphisms in *Bacillus anthracis*. *Science*, **296**, 2028–2033.

Reid, S.D., Herbelin, C.J., Bumbaugh, A.C., Selander, R.K., and Whittam, T.S. (2000) Parallel evolution of virulence in pathogenic *Escherichia coli*. *Nature*, **406**, 64-67.

Revazishvile, T., Rajanna, C., Bakanidze, L., *et al.* (2008) Characterisation of *Yersinia pestis* isolates from natural foci of plague in the Republic of Georgia, and their relationship to *Y. pestis* isolates from other countries. *Clinical Microbiology and Infection*, **14**, 429-436.

Richards, R. I., and Sutherland, G.R. (1997) Dynamic mutation: possible mechanisms and significance in human disease. *Trends in Biochemical Sciences*, **22**, 432–436.

U'Ren, J.M., Schupp, J.M., Pearson, T. *et al.* (2007) Tandem repeat regions within the *Burkholderia pseudomallei* genome and their application for high resolution genotyping. *BMC Microbiology*, **7**, 23.

van Belkum, A., Scherer, S., van Alphen, L., and Verbrugh, H (1998) Short sequence DNA repeats in prokaryotic genomes. *Microbiology and Molecular Biology Reviews*, **62**, 275–293.

Van Ert, M.N., Easterday, W.R., Huynh, L.Y., *et al.* (2007a) Global genetic population structure of *Bacillus anthracis*. *Plos One*, **5**, e461.

Van Ert, M.N., Easterday, W.R., Simonson, T.S., *et al.* (2007b) Strain-specific single-nucleotide polymorphism assays for the *Bacillus anthracis* Ames strain. *Journal of Clinical Microbiology*,

45, 47-53.

Vogler, A.M., Keys, C.E., Allender, C., *et al.* (2007) Mutations, mutation rates, and evolution at the hypervariable VNTR loci of *Yersinia pestis*. *Mutation Research*, **616**, 145-158.

Wattiau, P., Van Hesse, M., Neubauer, H., *et al.*, (2007) Identification of *Burkholderia pseudomallei* and related bacteria by multiple-locus sequence typing-derived PCR and real-time PCR. *Journal of Clinical Microbiology*, **45**, 1045-1048.

White, N.J. (2003) Melioidosis. *Lancet*, **361**, 1715-1722.

Zwick, M.E., Kiley, M.P., Stewart, A.C., Mateczun, A., and Read, T.D. (2008) Genotyping of *Bacillus cereus* strains by microarray-based resequencing. *PLoS ONE*, **3**, e2513.

Figure Legends

Figure 1. Identification of the *vrrA* allele in an unknown sample by direct comparison to a set of differentially labeled *vrrA* amplification fragments. VICTM-labeled *vrrA* PCR primers were used to amplify the *vrrA* allele from five different *B. anthracis* isolates, each containing a different *vrrA* allele. The resulting five allele fragments were purified, then mixed to make a molecular weight standard. DNA from an uncharacterized *B. anthracis* isolate was then used as template in a reaction containing FAMTM-labeled *vrrA* PCR primers. A small amount of the resulting amplicon was mixed with the VICTM-labeled *vrrA* standard and analyzed on an ABI 3100 capillary DNA sequencer. Numbers above each peak represent the number of (CAATATCAACAA) repeats present in the VNTR locus. The *vrrA* amplicon from the unknown samples migrates with the (CAATATCAACAA)₄ fragment. The slight size difference between the VICTM-labeled (CAATATCAACAA)₄ molecular weight control fragment and that from the unknown is due to differences in the molecular weights of the two dyes used to label the DNA fragments.

Figure 1.

